Normality of Errors

October 2, 2024

1 Normality of Residuals

Normality of Residuals is another important assumption in linear regression. It implies that the residuals should be normally distributed. In other words, the distribution of the residuals should be approximately normal. Mathematically, ϵ $N(0, \sigma^2)$ where σ^2 is the variance of the residuals, 0 refers to the mean, and $ $ $ is the residual. Thus, in linear regression, the assumption of normality means that for each combination of independent variable values, the residuals should be normally distributed with a mean of zero.

If the residuals are normally distributed then the coefficient estimates will be unbiased and have the minimum variance among all unbiased estimators. Also, the confidence intervals and p-values will be accurate, leading to valid inferences. However, it it noteworthy that the assumption of normality is not required for making predictions. If so then, the obvious question will be why normality of errors is so important?

1. The t-tests for individual regression coefficients (used to test whether the coefficients are significantly different from zero) and F-tests for overall model fit rely on the assumption that errors are normally distributed.
2. When errors are normally distributed, the confidence intervals for the regression coefficients are correctly specified. If the normality assumption is violated, confidence intervals might be too wide or too narrow.
3. Predictions made by the model, especially prediction intervals, assume that errors follow a normal distribution. If errors deviate significantly from normality, the prediction intervals may not capture the true uncertainty.

We can use various techniques to check for normality of errors. Some of the most common techniques are histogram of residuals, Q-Q Plot, Shapiro-Wilk Test, Kolmogorov-Smirnov Test etc.

Let's validate these concepts through codes.

[1]:

```
# import required libraries
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import seaborn as sns

from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

```
import statsmodels.api as sm
from scipy import stats
```

[2]:

```
# Prepare toy dataset
np.random.seed(0)
n = 100

# initialize random study hours between 0 and 10
X = np.random.rand(n, 1) * 10

true_coeff = 3.5
y = true_coeff * X.flatten() + np.random.normal(0, 5, size=n)
```

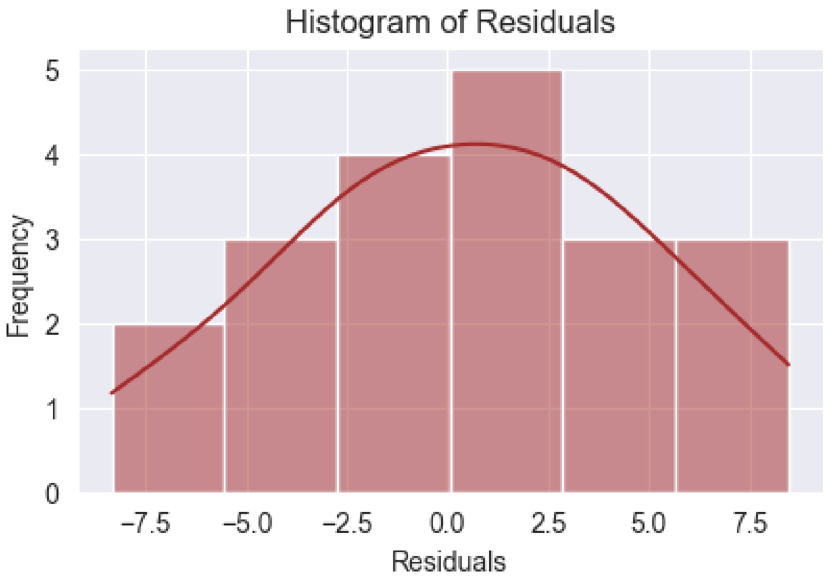
[3]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,␣
 ↪random_state=99)

# Fit linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# predict on test data
y_pred = model.predict(X_test)

# compute residuals
residuals = y_test - y_pred
```

1. Plot a histogram of residuals to visually inspect normality.

[4]:

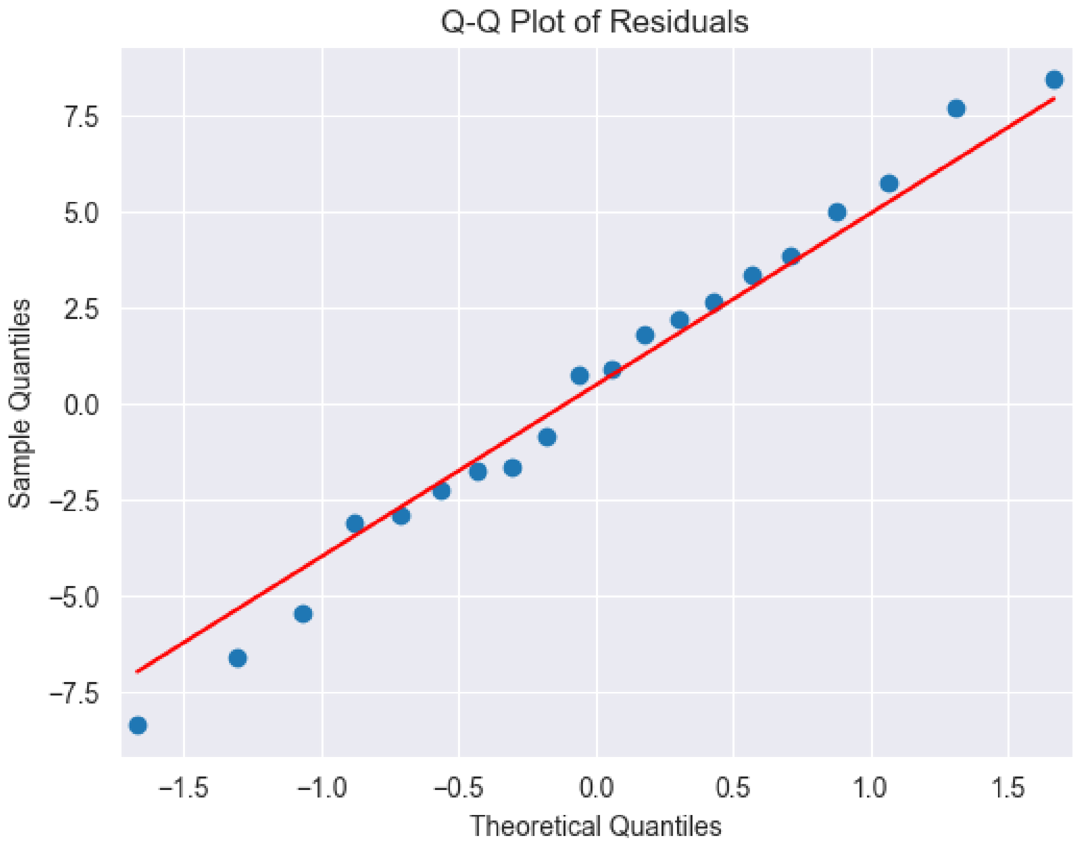
```
plt.figure(figsize=(5, 3))
sns.histplot(residuals, kde=True, color='brown')
plt.title('Histogram of Residuals')
plt.xlabel('Residuals')
plt.ylabel('Frequency')
plt.show()
```



2. Plot a Q-Q plot of residuals to visually inspect normality.

[5]:
```python
sm.qqplot(residuals, line='s', color='brown')
plt.title('Q-Q Plot of Residuals')
plt.show()
```

```
/Library/Frameworks/Python.framework/Versions/3.12/lib/python3.12/site-
packages/statsmodels/graphics/gofplots.py:1041: UserWarning: color is
redundantly defined by the 'color' keyword argument and the fmt string "b" (->
color=(0.0, 0.0, 1.0, 1)). The keyword argument will take precedence.
  ax.plot(x, y, fmt, **plot_style)
```



3. Perform Shapiro-Wilk Test for normality of residuals. Shapiro-Wilk Test for normality of residuals is a statistical test that checks whether a given dataset follows a normal distribution. It tests the null hypothesis that the data is normally distributed. It is one of the most commonly used tests for nom normality.

   Null Hypothesis (H): The data residuals follow a normal distribution.

   Alternative Hypothesis (H): The residuals do not follow a normal distribution.

The test statistic of this test is given as:

$$W = \frac{\sqrt{n}}{\sigma} \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

where, n is the number of observations, σ is the standard deviation of the residuals, and $\bar{x}$ is the mean of the residuals.

The p-value of Shapiro-Wilk test tells us the probability that the data could have been drawn from a normal distribution.

[6]:

```python
shapiro_statistic = stats.shapiro(residuals)
print("Shapiro-Wilk Test Results:")
print("Test Statistic:", shapiro_statistic.statistic)
print("p-value:", shapiro_statistic.pvalue)
```

```
Shapiro-Wilk Test Results:
Test Statistic: 0.9843237619760842
p-value: 0.9771394323190653
```

Interpretation of Shapiro-Wilk Test

1. P-value

- p > 0.05: Fail to reject H . This means the data appears to be normally distributed, as there is not enough evidence to conclude otherwise.
- p 0.05: Reject H . This indicates that the data does not follow a normal distribution.

In simple terms:

- If the p-value is greater than 0.05, we assume the residuals are normally distributed.
- the p-value is less than or equal to 0.05, we conclude that the residuals are not normally distributed.

2. Test Statistic (W):

- W 1: The closer the test statistic is to 1, the closer the data is to a normal distribution.
- W < 1: Values further from 1 indicate deviations from normality.

Interpreting the above results, the p-value is 0.9771, which is much larger than the conventional significance level of 0.05. This means we fail to reject the null hypothesis. In other words, there is no evidence to suggest that the data (or residuals) deviate significantly from a normal distribution. And, the test statistic, 0.9843, is very close to 1, which indicates that the distribution of the residuals is quite close to a normal distribution.

Thus, we can conclude that our regression model follows a normal distribution.

4. The Kolmogorov-Smirnov (K-S) test is another way to test whether a dataset follows a particular distribution, such as a normal distribution. Like the Shapiro-Wilk test, the K-S test compares the data to a reference distribution (in this case, the normal distribution) to assess normality. The Kolmogorov-Smirnov (K-S) test statistic measures the maximum difference between the empirical distribution function of the sample data and the cumulative distribution function of the reference distribution (in this case, the normal distribution). The K-S test statistic, denoted as D, is computed as:

$$D = \max |F_n(x) - F(x)|$$

Where:

- $F_n(x)$ is the empirical distribution function of the sample data.
- $F(x)$ is the cumulative distribution function of the reference distribution (normal distribution in this case).

The test statistic (D) reflects how much the distribution of the sample deviates from the normal distribution. A small value of D indicates that the sample is similar to the normal distribution, while a larger D value suggests that the sample distribution differs significantly from normality.

- $D = 0$: Perfect match with the normal distribution.
- $D > 0$: Bigger D values means greater deviation from the normal distribution.

The p-value associated with this test statistic helps determine if this deviation is statistically significant.

[7]:

```
from scipy import stats

# Kolmogorov-Smirnov test
ks_test = stats.kstest(residuals, 'norm', args=(np.mean(residuals), np.
  ↪std(residuals)))

# Print the test statistic and p-value
print(f"Kolmogorov-Smirnov Test Statistic: {ks_test.statistic}")
print(f"p-value: {ks_test.pvalue}")
```

```
Kolmogorov-Smirnov Test Statistic: 0.08272314439467632
p-value: 0.997266262225513
```

The KS test is also evaluated based on P-value as:

- If p > 0.05, we fail to reject the null hypothesis that the residuals follow a normal distribution.
- If p 0.05, we reject the null hypothesis and conclude that the residuals do not follow a normal distribution.

[8]:

```
if ks_test.pvalue > 0.05:
    print("Residuals are normally distributed (Fail to reject H0).")
else:
    print("Residuals are not normally distributed (Reject H0).")
```

```
Residuals are normally distributed (Fail to reject H0).
```